DOCUMENT RESUME

ED 267 869 JC 860 173

AUTHOR Belcher, Marcia
TITLE The Reliability of the CLAST. Research Report No. 84-19.
INSTITUTION Miami-Dade Community Coll., Fla. Office of Institutional Research.
PUB DATE Jul 84
NOTE 32p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Community Colleges; Criterion Referenced Tests; *Error of Measurement; *Ethnic Groups; *Minimum Competency Testing; Testing Problems; *Test Reliability; Test Results; Two Year Colleges; Two Year College Students
IDENTIFIERS *College Level Academic Skills Test

ABSTRACT

A study was conducted to assess the reliability of the College-Level Academic Skills Test (CLAST) for the major ethnic groups in Florida; and to develop a standard error of measurement for the critical group that fell close to the cut score on each test. The October 1983 and March 1984 administrations of the CLAST were analyzed to determine whether one ethnic group was consistently obtaining lower reliability estimates than other groups, the accuracy of each subtest in student placement, standard error, and subtest reliability. Study findings included the following: (1) the computation portion of the CLAST had the highest reliability, while the writing subtest had the lowest reliability; (2) the scores of Hispanics and Black non-Hispanics usually had the lowest reliabilities of any sub-group; (3) the size of the standard error of measurement around the cut score was somewhat higher than the traditional standard errors of measurement previously reported, indicating that students who fell below the cut score by 10-20 scale score points should be encouraged to retake the test. Information on previous reliability studies is included. (LAL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED267869

THE RELIABILITY OF THE CLAST

Research Report No. 84-19

July 1984

Marcia Belcher

Research Associate, Sr.

OFFICE OF INSTITUTIONAL RESEARCH

John Losak, Dean

MIAMI-DADE COMMUNITY COLLEGE

JC 860 173

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. McCabe

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Table of Contents

	Page
List of Tables	ii
Executive Summary	iii
Background on Reliability	1
Reliability of the CLAST	3
The Interinstitutional Research Council Study	4
DOE January 1984 Research Report	4
CLAST Technical Report 1982-83	6
Summary of Previous Reliability Studies	9
Purpose of the Study	10
Procedures	10
Results	12
Characteristics of the Group	12
Reliability of the Multiple Choice Tests	12
Standard Error of Measurement for Pass/Fail Multiple Choice Tests	16
Reliability of the Essay	20
Discussion	22
References	24

List of Tables

Tables		Page
1	IRC Reliability Results.	4
2	Index of Agreement	5
3	Reliability of Objective Subtests 1982-83 Administrations. .	7
4	Alpha Coefficients by Topic.	8
5	Performance on CLAST Subtests by Gender and Ethnic Membership	13
6	Reliability/Dependability Coefficients for the Multiple Choice Portion of the CLAST.	14
7	Split-Half Test Characteristics for Two Administrations Calculated for the Total Group	17
8	Relationship of Scaled to Raw Scores in the Area Around the Cutscore	18
9	Standard Error of Measurement Around the Cutscore.	19
10	Reliability of Essay Ratings Using Coefficient Alpha	21

Executive Summary

Reliability forms the base from which the worth of a test can be further judged. If stable scores cannot be obtained for people taking the test, then it is useless to ask further if the test provides a good measure of the construct it purports to measure.

Norm-referenced tests are designed to discriminate among people taking the test. Criterion-referenced tests are designed to measure test-takers against some standard of performance. Most traditional methods of measuring reliability were developed for norm-referenced tests. Other less well-known methods have been developed for criterion-referenced tests. Less agreement exists among measurement specialists on the approach which should be used to measure and interpret the reliability of criterion-referenced tests than on how to measure and interpret the reliability of norm-referenced tests.

The CLAST is a criterion-referenced test with enough items and score variability to allow traditional reliability approaches to be employed. In the past, both approaches have been used. It was found, for example, that for two administrations of the CLAST a high of 96% of the subjects were classified the same (pass/pass or fail/fail) on Computation while a low of 81% were consistently classified into pass/fail categories on the Essay. Using the traditional approach to reliability, it was found that close to 90% of the variability in scores on the Computation portion was due to "true" differences in computing abilities rather than to error. For the Writing portion, about 70% of the score variability was due to "true" writing ability differences on the Writing test. The standard error of measurement (SEM) indicated that if the test were given again, an individual's raw score could be expected to rise or fall by no more than two to three points 68% of the time (±1 SEM) and by no more than four to six points range (±2 SEM) 95% of the time, depending on the subtest being administered.

This study was designed to assess the reliability of the CLAST for the major ethnic groups and to develop a standard error of measurement for the critical group that fell close to the cut score on each test. Two types of

criterion-referenced indices, as well as the traditional internal consistency measures, were calculated. Questions to be answered included:

1) Is one group consistently obtaining lower reliability estimates than the other groups?
2) How accurate is each subtest in placing students whose scores fall close to the cut score?
3) Is the standard error small enough so that if they take test again, the same result probably will be obtained?
4) Does one subtest show consistently lower reliability than the others, no matter which coefficient is employed?

Results indicated that the Computation portion of the CLAST had the highest reliability, whether a norm-referenced or a criterion-referenced approach was used. The Writing subtest consistently had the lowest reliability, especially when traditional reliability coefficients were calculated. The main concern regarding the Reading subtest was not the size of the coefficients, but rather the number of students who failed to complete all items. Modification of the procedure for scoring essays to rescore any essay that was rated close to the cut seemed to result in increased reliabilities for this portion of the test.

Hispanics and black non-Hispanics usually had the lowest reliabilities for any sub-group. These low coefficients were accompanied by (and quite possibly were caused by) a smaller variability in the reported scores for the group. In terms of classical reliability theory, when small differences exist between members' scores, it is more difficult to discriminate on an ability scale as to who should be placed above whom; the result is more error within the group and decreased reliability coefficients.

The size of the standard error of measurement around the cut score was somewhat higher than the traditional standard errors of measurement previously reported. Results indicated that students who fell below the cut score by 10-20 scale score points should be encouraged to retake the portions they failed since by chance alone they had a good possibility of

scoring on the other side of the cut next time. Students slightly above the cut should be discouraged from retaking the CLAST without further remediation in their weak areas since they could just as easily fall below the cut as above it the next time the test was given.

Item bias and content analysis studies were suggested as other ways of studying the measurement properties of the CLAST. Since the CLAST could be consistent in classifying students as "masters" or "non-masters," yet inaccurate in identifying who should be so classified, it was urged that the placement of the cut scores should be revisited at the State level. At the local level, the impact of the curriculum on producing students who can pass the CLAST might be a good place to begin.

Background on Reliability

The first prerequisite of a good test is that it must be reliable (Mehrens & Lehmann, 1978). Reliability addresses the question of how consistently something is being measured by one or more tests. Reliability can be studied by giving the same test at two different points in time and looking for consistency in each person's score at the two times. It can be studied by giving each person two forms of a test, each of which is purported to measure the same thing, and checking for consistency in each person's set of scores. Another way of measuring reliability is to assess the internal consistency of the test to check that the test is measuring only one thing.

Several factors influence the reliability of the test. For example, the greater the number of items included in a test, the higher the reliability if the items all measure the same thing. In addition, the homogeneity of the group taking the test can influence the reliability. Other factors being equal, the more heterogeneous the group, the greater the variability in scores, and the higher the reliability of the test. Using the same reasoning, the difficulty of the test (and thus the individual items) affects the reliability because of the differences in score variance. If a test is very hard or very easy, little variability in scores will result and the reliability will decrease.

The variability of scores is an important component of classic reliability theory because the traditional concept of reliability is based on the assumption that the test is norm-referenced, i.e., designed to discriminate among individuals or compare them to one another. With criterion-referenced testing, where the object is to compare an individual against some external standard, the traditional concepts of reliability do not provide accurate information on the criterion-referenced test. For example, in a criterion-referenced test, it would be acceptable, and even desirable, for every member of the group to obtain a perfect score. Yet, using traditional concepts of reliability, low reliability coefficients would result.

Clearly, new methods of measuring reliability had to be developed which could accommodate the extremes of criterion-referenced tests and address the special issues of criterion-referenced testing, especially the consistency and precision of decisions on whether students had reached or exceeded the criterion and should be passed, or whether they had not and should be retained or failed. In response, more than a dozen statistics were devised (Berk, 1980). Most involved either the notion of consistency of mastery-nonmastery classification decisions across repeated measures or the reliability of the criterion-referenced test scores as measured by deviations from the cut score.

The concept of "classification decision consistency" was operationalized using several "threshold loss agreement indices" (e.g., Hambleton & Novick, 1973). These indices assumed that (1) students were classified as masters/nonmasters or pass/fail based on a threshold or cutting score, and (2) the losses associated with false mastery and false nonmastery classification errors were equally serious, whatever the size of the error.

The second concept of reliability was operationalized using "squared error loss agreement indices" (Hambleton et al., 1978). Instead of an either/or decision, these indices reflected a sensitivity to the degrees of mastery along a score continuum. Errors associated with misclassification were not considered to be equally serious. The larger the misclassification error, the greater the loss.

Measurement specialists have been unable to agree on one index to measure the reliability of criterion-referenced tests. Each index proposed has drawbacks in interpretation so that it cannot be directly compared to a classic reliability coefficient. In addition, the decision of whether to select a threshold loss agreement index or a squared error loss agreement index is at least in part a philosophical one. In selecting one type of index over the other, the consumer must look at the testing situation and agree with one or the other of the following statements:

A. To misclassify a student by one point is less serious than to misclassify a student by twenty points. The size of the error should be considered. If 10 students are failed by one or two points on Test A when they should have passed, while they are failed by 15-20 points on Test B when they should have passed, Test A should show a higher reliability than Test B because the errors of misclassification are smaller.

B. It doesn't matter whether a student is misclassified by one point or twenty points. The result is the same: failure. An error is an error.

People who select statement A will prefer squared error loss agreement indices. People who select statement B will prefer threshold loss agreement indices to assess reliability/consistency.

Reliability of the CLAST

Into this morass, let us now insert the issue of the reliability of the CLAST, a sophomore-level exit examination mandated by the legislature to "determine the extent to which college students have achieved the communication and computation skills expected of all students by the completion of their sophomore year" (College Level Academic Skills Project, 1983, p. 1). The Communications portion tests 35 specific skills, while Computation section tests 56 skills. The Technical Report of 1982-83 notes that the test is not "designed to yield skill-by-skill information needed for full diagnosis of the problems of individual examinees" (p. 2). Instead, responses are summed to yield a scaled score in Reading, Writing, Computation, and two readers rate an essay written by the student. Cutoff scores have been established for each portion of the CLAST. In order to receive an A.A. degree, the student must score above the cutscore on all four subtests. Since decisions are made on the basis of subtest scores, this, rather than specific skills, is the appropriate unit for reliability analysis.

Both classical and criterion-referenced measures have been employed to assess the reliability of the CLAST. The results along with comments on the drawbacks of each study are presented below.

The Interinstitutional Research Council Study

This unpublished study was conducted by the Institutional Research Council (IRC), an organization under the aegis of the Center for Higher Education at the University of Florida whose membership consists of two-year community colleges. A small sample (exact size not reported) of students whose scores fell below the mean or median were used to compute the split-half reliability for the CLAST Writing, Reading, and Computation subtests. The items were split using two methods: 1) an odd/even split with even items assigned to Test 1 and odd items assigned to Test 2; 2) a first/last split with the first half of the items assigned to Test 1 and the second half assigned to Test 2. The resulting coefficients are shown in Table 1 below. The author of the report, Dr. John Nickins, indicated that the results were similar for both methods of splitting the test (personal communication, March, 1984).

Table 1
IRC Reliability Results

Subtest	Split-Half	Corrected
Writing	.12	.21
Reading	.31	.47
Computation	.58	.73

This study has several problems. The small sample size reduced the stability of the correlation. The range of scores was restricted, which lowered the correlation. The correlation was further lowered when the halves of the test were not matched for content. The use of split-half correlation resulted in a lower correlation because only half of the items were used to calculate the reliability. When corrected using the Spearman-Brown prophecy formula, the results changed as shown in the second column of Table 1.

DOE January 1984 Research Report

In this study two criterion-referenced methods were used to calculate the reliability of the CLAST. In one method the Brennan and Kane

(1977) Dependability Index was computed on the October 1983 performance data for over 14,000 students. This index is based on squared deviations of individual scores around the cutting score and requires only one administration. The assumption behind this procedure is that "near misses" in correctly classifying a student as a master or nonmaster are not as serious as larger measurement errors. The results for the study were:

Reading	.97
Writing	.92
Computation	.96

In a second reliability study, the threshold loss approach (Hambleton & Novick, 1973) was employed. A small sample of students (n=97) was paid to take a second version of the CLAST approximately one week after the June 1983 administration. For each subtest, the results were cast into a table like that shown below.

Test 1

Test 2		Pass	Fail
	Fail	P_{fp}	P_{ff}
	Pass	P_{pp}	P_{pf}

$P_o = P_{pp} + P_{ff}$

The proportions of consistent decisions (pass-pass, fail-fail) were summed for each subtest to form an Index of Agreement, P_o. The results are shown in the first column of Table 2 below.

Table 2

Index of Agreement

Subtest	P_o	Kappa
Reading	.91	.52
Writing	.92	.30
Computation	.96	.69
Essay	.81	.31

This second study was generally handicapped by a small sample size and a possibly unrepresentative sample since the students were poorly paid volunteers. Using Kappa (K), which corrects for chance, considerably decreased P_o. A problem with this statistic, however, is that Kappa is equal to 1.0 only when the proportions passing each test are equal (e.g., 70% pass the test on each occasion). Table 2 lists the Kappa coefficients.

Both indices in the DOE study are higher since they are uncorrected for chance agreement. Both indices increase as:

1) the cutscore moves further from the mean;
2) the number of items increases; and
3) score variance increases.

The CLAST has: 1) a cutscore which falls between 35 and 40 scale score points below the mean, 2) a large number of items, and 3) comparatively large score variance for a criterion-referenced test.

CLAST Technical Report 1982-83

Typically, when traditional reliability indices are used on criterion-referenced tests, the coefficients obtained are low. This finding occurs because coefficients such as KR-20 and coefficient alpha increase as the number of items and score variance increase, and most criterion-referenced tests are characterized by few items and little score variance. This is not true of the CLAST, so traditional measures of reliability are more appropriate than they might be for many criterion-referenced tests.

For the Reading, Writing, and Computation subtests, internal consistency reliability was assessed by calculating KR-20 on the total group who took the October 1982, March and June 1983 administrations. The results, along with the standard error of measurement (SEM), can be found in Table 3 below.

Table 3

Reliability of Objective Subtests
1982-83 Administrations

	Reading			Writing			Computation		
	Oct.	March	June	Oct.	March	June	Oct.	March	June
KR-20	.87	.85	.85	.72	.68	.69	.90	.88	.88
SEM	2.29	2.19	2.25	1.81	1.81	1.85	2.84	2.98	3.01

The KR-20 coefficients can be interpreted to mean that between 85-87% of the variability in Reading scores is due to "true" differences between the people taking the test, while only 68-72% of the variability in Writing scores reflects these differences. Therefore, the Writing test has more measurement error than does the Reading test; Computation has the least measurement error in the scores.

The standard error of measurement (SEM) provides an indication of how much a persons's score could be expected to vary if the test, or one like it, were given again. For example, in the June administration, the Computation SEM was about 3. Therefore, if the Computation test were given again, we could expect a person's estimated ability score to fall within three points above or three points below the first score 68% of the time (±1 SEM) and within six points above or below 95% of the time (±2 SEM). If someone obtained a score of 40, we would predict that if that person took the Computation test again, he/she would probably obtain a score between 34 and 46 on the next administration, assuming additional learning or other factors did not interfere between the two testings to alter the person's "true" level of computational ability as measured by the test.

The Essay ratings' reliability was assessed in two ways: 1) reader agreement on assignment of scores and 2) coefficient alpha. The raters were in total agreement between 51 and 53 percent of the time; they were within one point of one another or in total agreement 96% of the time. For two topics, coefficient alpha for the total group ranged from a low of .76 to a high of .82, indicating that 76-82% of the variability in ratings was due to "true" differences among the essays. A further breakdown can be found in Table 4 (copied from the Technical Manual).

Table 4

Alpha Coefficients by Topic

	Topic One With Referee			Topic Two With Referee		
	October 1982	March 1983	June 1983	October 1982	March 1983	June 1983
All Students	79	76	80	80	82	79
Males	79	76	80	81	81	78
Females	79	75	78	79	82	79
Whites	75	72	76	76	77	74
Blacks	81	76	80	80	83	75
Hispanics	82	77	80	82	86	80
Indian/Alaskan	85	84	96	81	83	79
Asian	84	83	87	90	81	85
CC-AA[1]	80	77	80	80	82	79
CC-AS[2]	81	78	84	81	84	78
University Native		79	77		84	76
University Transfer		72	80		78	80

[1] Community college Associate of Arts students
[2] Community college Associate of Science students

Coefficient alpha and KR-20 set an upper limit to the reliability that can be attained with an instrument. Nunnally (1978) stated that: "In those applied settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated and a reliability of .95 should be considered the desirable standard" (p. 246). Most coefficients failed to reach this standard.

The size of these correlations is worrisome for another reason. The larger the correlation, the greater the indication that the test is measuring only one thing (i.e., it is unidimensional). The CLAST was standardized using the RASCH model. Use of the RASCH model requires that the test be unidimensional. If it is not, additional error and unreliability are introduced.

Summary of Previous Reliability Studies

Using two criterion-referenced methods, the reliability of the CLAST appears high. Except for Essay, all coefficients were in the 90's. Using a third measure, however, which corrected for chance but would only reach its maximum when the same proportions had failed each of the two administrations, significantly lowered the results. A good example was Reading, where the proportions passing each administration differed only by .01, yet the index of agreement dropped from .91 to .52. Using KR-20 and coefficient alpha, two traditional methods of assessing the upper limits of reliability, acceptable results were obtained for Computation and perhaps Reading, but unacceptable coefficients were found for Writing and for the Essay.

Except for the Essay portion, no attempt was made to see if the reliability was unusually low for a particular ethnic group or gender. The study by the IRC attempted to look at the reliability of CLAST for those negatively impacted by the test, but several measurement issues clouded the interpretability of their findings.

Purpose of the Study

This study was designed to assess the reliability of the CLAST for the major ethnic groups and to develop a standard error of measurement for the critical group that fell close to the cutscore on each test. Both threshold loss and squared error loss indices, as well as the traditional internal consistency measures, were calculated.

Questions to be answered included:

- Is one group consistently obtaining lower reliability estimates than the other groups?

- How accurate is each subtest in placing students whose scores fall close to the cutscore? Is the standard error small enough so that if they take the test again, the same result probably will be obtained?

- Is one subtest consistently lower than the others, no matter which coefficient is employed?

Procedures

The October, 1983, and March, 1984, administrations of the CLAST were used for the analysis. Data included the responses from all 1,561 students from the October test and 1,205 from the March test. The SPSS statistical programs were used to perform much of the data analysis.

To assess the reliability of the multiple choice portions of the CLAST, the following coefficients were calculated for the total group, each gender, and for three major ethnic groups (white non-Hispanics, black non-Hispanics, Hispanics):

(1) Internal consistency as measured by KR-20 and KR-21;
(2) Dependability Index (Brennan and Kane, 1977);

(3) Index of Agreement (P_o) and Kappa (K) using the single-administration approach advocated by Peng and Subkoviak (1980).

Using these coefficients allowed comparison of traditional approaches (1) to a squared error loss approach (2) to two threshold loss approaches (3).

To calculate the standard error measurement (SEM) for tests used to make pass-fail decisions, an approach suggested by Livingston (1982) was employed. Calculation of Livington's SEM involved:

- splitting the test in halves, matching for content and difficulty of the two halves;
- selecting from the total group all students whose scores fell close to the cutscore;
- calculating each student's score for each of the two halves;
- finding the difference between the two halves, squaring that difference, dividing by the number of students, then taking the root to obtain the SEM. This step can be symbolized as:

$$SEM = \sqrt{\frac{\Sigma (X_{1i} - X_{2i})^2}{n}}$$

Assessing the reliability of the Essay portion of the CLAST required a somewhat different approach since the Essay score was the result of two ratings rather than of a series of items. A traditional coefficient is coefficient alpha, the same method as KR-20 except the formula was modified so scores did not have to be based on either passing or failing a series of items. Alpha was calculated both when scores were refereed and when they were not. The frequency with which the raters agreed was also noted. All calculations were based on the total group, then recalculated based on gender and ethnic membership.

Results

Characteristics of the Group

Table 5 contains the summary statistics relevant to the analysis. Based on the table and prior knowledge of reliability theory, if all else were equal, we would predict that:

*the highest reliability coefficients should be found for Computation since it had the greatest number of items;

*the lowest traditional reliability coefficients should be found for Hispanics since they typically had the least variance in their scores;

*the lowest Dependability coefficients (Brennan & Kane, 1977) and Index of Agreement (P_o) coefficients should be found for black non-Hispanics since the mean score for this group was closer to the cutscore than it was for any other group;

*the highest Kappa coefficients should be found for black non-Hispanics for the same reason just cited.

Table 5 also shows the proportion of students from each subgroup who were included in the analysis. This statistic is important since if a student failed to respond to one or more items on a subtest, all responses by that student were excluded from the analysis. Note that black non-Hispanic students were most likely to leave items blank, so a smaller proportion of this subgroup was included in the analysis. This fact, combined with the already relatively small number of students belonging to this category, led to the conclusion that results for black non-Hispanics probably would not be as stable as results for groups with a greater number of members.

Table 5

Performance on CLAST Subtests by Gender and Ethnic Memberships

	October 1983				March 1984			
	Mean	S.D.	Percent Included in Analysis	Percent Passing	Mean	S.D.	Percent Included in Analysis	Percent Passing
Reading			(35 Items)				(36 Items)	
Total	25.6	5.8	.83	.94	24.8	5.9	.83	.87
Males	25.0	6.0	.84	.93	24.6	5.9	.83	.87
Females	26.1	5.6	.82	.95	25.1	5.9	.83	.87
White	27.5	5.3	.87	.96	27.3	5.0	.86	.96
Black	21.8	6.8	.70	.86	20.2	6.6	.71	.65
Hispanic	25.3	5.3	.85	.95	24.1	5.6	.84	.87
Writing			(35 Items)				(35 Items)	
Total	27.2	4.7	.96	.83	26.4	4.5	.94	.88
Males	26.8	4.7	.97	.82	26.1	4.4	.94	.86
Females	27.5	4.6	.96	.84	27.3	4.4	.94	.89
White	28.7	4.5	.97	.91	28.5	3.8	.96	.95
Black	24.8	5.2	.94	.63	24.4	4.6	.85	.79
Hispanic	26.9	4.3	.97	.83	26.2	4.4	.94	.86
Computation			(48 Items)				(54 Items)	
Total	31.5	8.7	.87	.90	35.5	8.7	.83	.94
Males	32.8	8.6	.88	.91	36.3	8.7	.83	.95
Females	30.2	8.7	.87	.88	34.7	8.7	.83	.93
White	33.4	8.6	.88	.93	37.6	8.6	.86	.97
Black	26.0	9.4	.80	.74	29.0	8.9	.71	.81
Hispanic	31.5	8.2	.90	.91	35.1	8.3	.84	.95
Essay								
Total	4.1	1.4	.99	.65	4.2	1.5	1.00	.76
Males	3.8	1.4	1.00	.60	3.9	1.4	.99	.72
Females	4.3	1.4	.99	.70	4.5	1.5	1.00	.80
White	4.5	1.4	1.00	.77	4.8	1.4	1.00	.88
Black	3.4	1.3	.99	.47	3.8	1.6	1.00	.63
Hispanic	4.0	1.3	1.00	.64	4.0	1.4	1.00	.73

Table 6

Reliability/Dependability Coefficients for the Multiple Choice Portion of the CLAST

October 1983 Administration					
Reading	**KR20**	**KR21**	**Mc**	**Po**	**Kappa**
Total	.84	.82	.96	.93	.55
Male	.84	.83	.95	.92	.57
Female	.83	.81	.96	.94	.54
White	.83	.82	.97	.96	.49
Black	.86	.85	.92	.88	.63
Hispanic	.80	.77	.95	.93	.48
Writing					
Total	.77	.74	.88	.86	.53
Male	.77	.74	.87	.90	.53
Female	.77	.74	.89	.87	.52
White	.79	.77	.93	.90	.51
Black	.78	.76	.81	.82	.57
Hispanic	.72	.68	.86	.85	.44
Computation					
Total	.89	.88	.96	.93	.66
Male	.89	.88	.97	.98	.92
Female	.89	.87	.97	.94	.72
White	.89	.88	.97	.94	.64
Black	.90	.88	.93	.89	.71
Hispanic	.88	.86	.96	.94	.62
March 1984 Administration					
Reading	**KR20**	**KR21**	**Mc**	**Po**	**Kappa**
Total	.82	.80	.97	.89	.56
Male	.81	.79	.92	.89	.55
Female	.82	.80	.93	.96	.56
White	.77	.75	.95	.94	.44
Black	.84	.82	.86	.82	.61
Hispanic	.79	.76	.91	.89	.55
Writing					
Total	.74	.70	.90	.89	.49
Male	.72	.67	.89	.87	.45
Female	.74	.70	.92	.90	.49
White	.69	.64	.94	.93	.39
Black	.71	.66	.83	.83	.47
Hispanic	.71	.67	.89	.87	.36
Computation					
Total	.87	.86	.96	.97	.78
Male	.87	.86	.97	.95	.55
Female	.87	.85	.96	.94	.61
White	.87	.86	.97	.97	.59
Black	.86	.85	.92	.88	.63
Hispanic	.86	.84	.96	.95	.54

Reliability of the Multiple Choice Tests

The results of the five selected indices are displayed in Table 6. For traditional measures of reliability (KR-20 and KR-21), the highest coefficients were obtained for Computation while the lowest were obtained for Writing. The coefficients were slightly lower for the March than for the October administration. In general, the results for the total group are close to the coefficients reported in the 1982-83 Technical Report.

When comparing the results by subgroup for each test, on Computation the coefficients changed very little. On the October Reading and Writing subtests, Hispanics scored somewhat lower than the other groups. In March, the white non-Hispanics were lowest. In each case, the group with the lowest reliability had the smallest standard deviation (see Table 1). This finding can be interpreted as indicating that subgroups such as Hispanics who show a smaller range of scores will have more error variance as part of the total score variance than will groups who display scores over a consistently wider range. The rationale is that it is more difficult to discriminate among people with scores which are close together than when scores are farther apart.

The Dependability Indices (Mc) were consistently higher than KR-20 and KR-21. This fact is a function of the formula since only if the group mean equaled the cutscore would the Dependability Index be as low as KR-20. Computation once again had the highest coefficients, closely followed by Reading. Writing had the lowest. By subgroup, the lowest results were consistently obtained for black non-Hispanics, whose mean was closest to the cutscore of any group. While Reading and Computation results paralleled those from the DOE research report, Writing results were slightly lower in this study.

The pattern of results for the Index of Agreement (P_o) mimicked that of the Dependability Index. The index indicated, for example, that 82% of black non-Hispanics would be similarly classified on two administrations of the Writing subtest, while 90% of white non-Hispanics would be.

Compared to the DOE study, the Writing Agreement Index was decidedly lower. Reading and Computation results were about the same as previously reported.

Kappa, the threshold loss index which corrects for chance, presented a different pattern of results. In part, this finding occurred because as the mean approaches the cutscore, Kappa gets higher, not lower as for Mc and P_o. Therefore, black non-Hispanics now obtained the highest coefficients. The Kappa coefficients generally were lower both because of the chance correction and because Kappa is capable of reaching the value of 1.0 only under very constrained circumstances.

Standard Error of Measurement for Pass/Fail Multiple Choice Tests

Recall that the first step in calculating Livingston's SEM was to divide the test into two halves which had been matched for content and difficulty. Table 7 displays the results of the split when the total group was employed in the analysis. The means and standard deviations were quite similar for the two tests. The correlations between the halves, when corrected for the total number of items, were similar to the results found using KR-20 and KR-21. The traditional method of calculating the standard error of measurement and Livingston's procedure also produced similar results. The only difference between this SEM and the one reported (see bottom of Table 7) in the Technical Manual is that the Writing SEM is .5 larger than previously reported.

The second step was to select all students whose scores for each subtest fell close to the cutscore. Table 8 displays the scaled scores, the number of correct items needed to reach that scaled score, and the number of students who obtained each score. Note that 10% of the students scored within one or two items of the cutscore in Writing and as few as 3% to 5% were that close to the cut in Computation. To conduct the analysis, only those students who obtained the raw score closest to the cut were included. Therefore, all students selected for the SEM analysis had the same total score on that subtest.

Table 7

Split-Half Test Characteristics
for Two Administrations
Calculated for the Total Group

	October 1983			March 1984		
	Reading	Writing	Computation	Reading	Writing	Computation
Mean of						
Test 1	12.1	13.1	16.5	12.3	12.5	17.0
Test 2	12.5	13.2	14.9	12.3	12.8	18.5
Standard Deviation						
Test 1	3.1	2.5	4.6	3.3	2.6	4.6
Test 2	3.1	2.5	4.6	3.2	2.4	4.6
Split-Half	.73	.64	.82	.66	.59	.80
Corrected	.84	.78	.90	.80	.74	.89
SEM	2.3	2.3	2.9	2.5	2.3	3.1
SEM-L	2.3	2.1	3.2	2.7	2.3	3.3

Table 8

Relationship of Scaled to Raw Scores
in the Area Around the Cutscore

	Scaled Score	Raw Score	Number Obtaining Score	Percent of Total Group
October 1983 Administration				
Reading:	256	14	23	1%
(cut = 260)	260	15*	52	3%
	264	16	28	2%
Subtotal:			103	7%
Writing:				
(cut = 265)	259	21	50	3%
	264	22*	56	4%
	268	23	67	4%
Subtotal:			173	11%
Computation:				
(cut = 260)	256	18	30	2%
	259	19*	22	1%
	262	20	33	2%
Subtotal:			85	5%
March 1984 Administration				
Reading:				
(cut = 260)	257	16	28	2%
	261	17*	37	3%
	264	18	41	3%
Subtotal:			106	9%
Writing:				
(cut = 265)	260	19	33	3%
	264	20*	41	3%
	269	21	48	4%
Subtotal:			122	10%
Computation:				
(cut = 260)	258	20	10	1%
	261	21*	18	1%
	263	22	8	1%
Subtotal:			36	3%

*Selected for further analysis.

Table 9

Standard Error of Measurement
Around the Cutscore

	Mean of Test 1	Mean of Test 2	Standard Deviation of Test 1	Standard Deviation of Test 2	Number in Group	Livingston's S.E.M.
Reading:						
October	7.2	7.2	1.2	1.1	38	2.3
March	9.3	8.1	1.6	1.7	25	3.4
Writing:						
October	10.5	10.9	1.4	1.3	55	2.6
March	9.1	10.1	1.2	1.2	37	2.5
Computation:						
October	10.4	8.3	1.7	1.8	19	4.0
March	10.2	10.6	1.8	1.9	15	3.6

Table 9 shows the results of applying the two test halves to the groups and obtaining the standard errors of measurement for each subtest. The SEMs for this group were all slightly larger than they were for the total group. These SEMs can be interpreted as the average number of items students' scores can be expected to change over repeated testing. The formula takes into account the possibility that one-half of the test may consistently be more difficult than the other half. On the Writing subtest, for example, 68% of the time we could expect scores to vary by at least 2.5 items (or about 10 scaled score points) for students around the cutscore.

Reliability of the Essay

The reliability of the Essay portion of the CLAST hinges on the agreement of the raters who read the essays. In October, 1983, 56% of the time both raters agreed on M-DCC students' essay scores; 97% of the time the raters were within one point of one another. In March similar results were obtained. The raters agreed on the score for an essay 55% of the time; 96% of the time they were within one point of one another.

Coefficient alpha also assesses the agreement among raters. Table 10 contains the results when the raters had individually given their ratings on the essays (not refereed) and again after questionable essays had been reviewed and rescored by a referee. Note that the refereeing process improved the reliability of the scores. The March, 1984, reliability coefficients were higher than October's probably because starting with the March administration, any score which received a "3" was rescored as either a "2" or a "4."

Table 10

Reliability of Essay Ratings Using Coefficient Alpha

	October 1983			March 1984		
	Refereed	Not Refereed	Number	Refereed	Not Refereed	Number
Total	.79	.72	1,561	.89	.73	1,205
Males	.80	.72	738	.89	.70	545
Females	.78	.70	823	.88	.74	660
White	.78	.68	488	.83	.70	392
Black	.81	.74	202	.93	.81	130
Hispanic	.77	.69	839	.89	.69	662

Discussion

The purpose of this study was to assess the reliability of the CLAST and to compare the results for the four subtests by gender and ethnic membership. The reliability coefficients, whether developed for traditional norm-referenced measures or specifically for criterion-referenced measures, indicated that the Computation subtest was the most reliable and Writing was the least reliable. Hispanics and black non-Hispanics usually had the lowest reliability coefficients, though most differences among the subgroups could be traced to the interaction of response characteristics of the group and the formula used to calculate the coefficients.

Are the obtained coefficients "high enough" for a test used to make important decisions about students? For Computation, the answer is a qualified "yes." The answer is qualified because while higher is better, too many additional items probably would be needed to boost the coefficients. With the change in scoring the Essay, reliability is sufficient, and it is unlikely that the Essay reliability could be improved much more. The Reading and Writing subtests' reliability are at less desirable levels. The coefficients obtained for the Writing subtests are particularly low. While the Reading coefficients are higher, it appears that numerous students are not finishing the test, perhaps because of time constraints.

The size of the standard errors of measurement provided an indication that the CLAST does not discriminate finely enough at the cutting score to make an accurate assessment of who should pass and who should fail. As a result, a student who scores within 10 to 20 scaled score points of the cutscore should be counseled to take the test again since on the next testing his/her score could easily exceed the cutscore, based on chance factors alone. Students who score slightly above the cutscore, of course, should leave well enough alone.

Future studies should focus on several issues related to the reliability of the CLAST. One issue is item bias. Students from a similar cultural background may score consistently on a test (i.e., reliably), but

the test may be measuring something different from what is being measured for other cultural groups. A second issue is the number of constructs the test is measuring. The test was designed to measure communication and computation skills. The question is whether only these two aspects are being measured and if the items which measure them have been properly grouped together. The idea behind this type of analysis is that, for example, a computation item should depend more on computation skills than on reading ability. The greater the number of constructs being measured by the test, the more error introduced into the scaling process and into the process used to equate one form of the CLAST to the next administration.

This study addressed the reliability of the CLAST from several different viewpoints. In some ways, however, the study has placed the proverbial cart before the horse. As Berk (1980) noted,

> Without validity evidence or a sound justification for setting the cutting score, it seems pointless even to compute an agreement index... . A high agreement index associated with an "invalid" or "unjustified" standard, for example, might indicate that a test can consistently classify students into the wrong groups. Consistent decision-making without accurate decision-making has questionable value in criterion-referenced evaluation (p. 325).

Who are the true "masters" and "nonmasters" of the computation and communication skills identified by the legislature and where do they score on the CLAST? What should these "masters" be able to do that "nonmasters" cannot as a result of their mastery of the identified skills? What loss should be associated with falsely identifying a student as a "nonmaster" vs. the loss of allowing a student to go on because of falsely identifying that student as a "master"? This study has worked around these issues. Eventually, however, the basic issue of establishing a valid cutscore must be readdressed at the State level. At the local level, we could begin to assess the impact of the curriculum in turning "non-masters" into "masters."

REFERENCES

Berk, R. A. (1980). A consumer's guide to criterion-referenced test reliability. Journal of Educational Measurement, 17, 323-349.

Brennan, R. L. & Kane, M. T. (1977). An index of dependability for mastery tests. Journal of Educational Measurement, 14, 277-289.

College Level Academic Skills Project. (1983). College Level Academic Skills Test Technical Report 1982-83. Tallahassee, Florida: Department of Education.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criteron-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Livingston, S. A. (1982). Estimation of the conditional standard error of measurement for stratified tests. Journal of Educational Measurement, 19, 135-139.

Mehrens, R. K., & Lehmann, I. J. (1978). Measurement and evaluation in education and guidance (2nd ed). New York: Holt, Rinehart, and Winston.

Nunnally, J. C. (1978). Psychometric theory (2nd ed). New York: McGraw-Hill Book Company.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. Journal of Educational Measurement, 17, 259-268.

State of Florida Department of Education. (1984). The reliability of classification decisions on CLAST (Research Report 01-84). Tallahassee: Department of Education.

ERIC CLEARINGHOUSE
FOR JUNIOR COLLEGES
MAY 15 1986